University of Zurich[UZH]

Zurich Open Repository and Archive
University of Zurich
University Library
Strickhofstrasse 39
CH-8057 Zurich
www.zora.uzh.ch

Year: 2007

The neural signature of social norm compliance

Spitzer, M ; Fischbacher, U ; Herrnberger, B ; Grön, G ; Fehr, Ernst

Abstract: All known human societies establish social order by punishing violators of social norms. However, little is known about how the human brain processes the punishment threat associated with norm violations. We use fMRI to study the neural circuitry behind forced norm compliance by comparing a treatment in which norm violations can be punished with a control treatment in which punishment is impossible. Individuals' increase in norm compliance when punishment is possible exhibits a strong positive correlation with activations in the lateral orbitofrontal cortex and right dorsolateral prefrontal cortex. These activations are also modulated by the social nature of the task. Moreover, activation in lateral orbitofrontal cortex shows a strong positive correlation with Machiavellian personality characteristics. These findings indicate a neural network involved in forced norm compliance that may constitute an important basis for human sociality. Different activations of this network reveal individual differences in the behavioral response to the punishment threat and may thus provide a deeper understanding of the neurobiological sources of pathologies such as antisocial personality disorder.

DOI: https://doi.org/10.1016/j.neuron.2007.09.011

Other titles: The neural signature of forced norm compliance (running title)

Posted at the Zurich Open Repository and Archive, University of Zurich
ZORA URL: https://doi.org/10.5167/uzh-3834
Journal Article
Accepted Version

Originally published at:
Spitzer, M; Fischbacher, U; Herrnberger, B; Grön, G; Fehr, Ernst (2007). The neural signature of social norm compliance. Neuron, 56(1):185-196.
DOI: https://doi.org/10.1016/j.neuron.2007.09.011

Published in: Neuron 2007

The Neural Signature of Forced Norm Compliance

Manfred Spitzer[1,2], Urs Fischbacher[3], Bärbel Herrnberger[1], Georg Grön[1], Ernst Fehr[3,4]

[1] University of Ulm, University Hospital for Psychiatry, Psychiatry III, Leimgrubenweg 1214, 89075 Ulm, Germany

[2] Transfer Center for Neurosciences and Learning (ZNL), Beim Alten Fritz 2, 89075 Ulm, Germany

[3] Institute for Empirical Research in Economics, University of Zurich, Blümlisalpstrasse 10, 8006 Zurich, Switzerland

[4] Collegium Helveticum, Schmelzbergstrasse 25, 8092 Zürich, Switzerland

Running title: Forced Norm Compliance

Corresponding authors:

Manfred Spitzer, Ph. D.
Dept. of Psychiatry

University of Ulm
Leimgrubenweg 12
89075 Ulm
Germany
Tel.: +49 731 500 6
Fax: +49 731 500
e-mail: manfred.spitzer@uni-ulm.de

Ernst Fehr
Institute for Empirical Research in Economics
University of Zurich
Blümlisalpstrasse 10
8006 Zürich
Switzerland
Tel.: +41 44 634 37 09
Fax: +41 44 634 49 07
e-mail: efehr@iew.uzh.ch

Summary

All known human societies establish social order by punishing violators of social norms. However, little is know about how the human brain processes the punishment threat associated with norm violations. We use fMRI to study the neural circuitry behind forced norm compliance by comparing a treatment in which norm violations can be punished with a control treatment in which punishment is impossible. Individuals' increase in norm compliance when punishment is possible exhibits a strong positive correlation with activations in the lateral orbitofrontal cortex and right dorsolateral prefrontal cortex. These activations are also modulated by the social nature of the task. Moreover, activation in lateral orbitofrontal cortex shows a strong positive correlation with Machiavellian personality characteristics. These findings indicate a neural network involved in forced norm compliance that may constitute an important basis for human sociality. Different activations of this network reveal individual differences in the behavioral response to the punishment threat and may thus provide a deeper understanding of the neurobiological sources of pathologies such as antisocial personality disorder.

Introduction

Humans are unique among all species in the extent to which they regulate social life through compliance with social norms. Such norms constitute standards of behavior that are based on widely shared beliefs on how individuals ought to behave in a given situation (Elster, 1989). Ethnographic evidence (Sober and Wilson, 1998), evolutionary theory (Boyd et al., 2003), and laboratory studies (Fehr and Gächter, 2002) indicate that the maintenance of social norms typically requires a punishment threat, as there are almost always some individuals whose self-interest tempts them to violate the norm. Norm obedience sometimes vanishes quickly in the absence of a credible punishment threat. Following the death of Valentinian III in 455 A.D., the vandals invaded and looted Rome. Police strikes in Liverpool in 1919 and in Montreal in 1956 caused huge increases in crime rates (Adenaes, 1974), and during the terrorist attacks of September 11, 2001 – when the police was occupied with responding to the disaster – many ATMs were robbed in New York. The dissolution of obedience to prevailing norms occurs because people often comply with social norms conditional on others' compliance (Fischbacher et al., 2001). Thus, even a minority of non-compliers can trigger a process that induces widespread defection from prevailing norms.

Legal institutions such as the police and the courts enforce many norms in most contemporary human societies. However, legal enforcement mechanisms cannot function unless they are based on a broad consensus about the normative legitimacy of rules, that is, unless social norms support legal rules. The very existence of legal enforcement institutions is itself a product of prior norms about what constitutes appropriate behavior. Private sanctions have enforced social norms for millennia, long before legal enforcement institutions existed, and punishment by peers still represents a powerful norm enforcement device, even in contemporary Western societies (Fehr and Gächter, 2002). In view of the prominent role of such peer punishment in human evolution (Sober and Wilson, 1998; Boyd et al., 2003),

humans have developed elaborate neural mechanisms for social cognition that produce appropriate responses to the threat of peer punishment. However, although important work examining the neural basis of economic choice (Glimcher et al., 2005; Hsu et al., 2005), social cognition (Adolphs, 2001, 2003), moral judgment (Greene et al., 2001; Greene et al., 2004; Moll et al., 2002; Moll et al., 2005; Raine and Yang 2006), social cooperation (Rilling et al., 2002; Singer et al. 2006; Delgado et al, 2005; Kings Casas et al., 2005), and social punishment (Sanfey et al., 2003; DeQuervain et al., 2004; Knoch et al. 2006) exists, the brain systems involved in forced norm compliance still remain unknown. In particular, the previous literature on social punishment examines the neural circuitry involved in the decision to punish whereas our work focuses on the neural circuitry involved in the processing and the response to punishment threats that are associated with norm violations.

In the present study, we combined a behavioral experiment, involving real monetary stakes and the requirement to curb immediate self-interest in order to obey a fairness norm, with functional magnetic resonance imaging (fMRI) to study how the threat of punishment enforced norm compliance in subjects who have to decide whether to obey or violate the fairness norm (Kahnemann et al., 1986). If subjects violate the norm in the control condition they face no sanctions whatsoever. In the punishment condition, however, the victim of the norm violation could punish the perpetrator. As violations of fairness norms motivate the victim to punish the violator (Fehr and Fischbacher, 2004; Fehr and Gächter, 2002; Güth et al., 1982; Kahnemann et al., 1986), potential norm violators were to face a serious punishment threat in the punishment condition.

Two players, A (in the scanner) and B, interacted anonymously with each other (see also Experimental Procedure). Each knew that he was facing a human player. Player A received an endowment of 100 money units (MUs) which he could distribute freely between himself and player B. In the control condition, which resembles a dictator game, B was a

passive recipient (Kahneman et al., 1986) of A's monetary transfer. In contrast, B could punish A in the punishment condition after having been informed of the latter's decision (Andreoni et al., 2003). Each player received an additional endowment of 25 MUs in both conditions, for reasons of fairness and to make punishment possible. In the punishment condition, B could spend all or part of this amount to reduce A's earnings; every MU B invested into punishment led to a reduction of A's earnings by 5 MUs. For example, if A kept all the 100 MUs for himself and B punished maximally, then A's earnings (100 MUs initial endowment plus 25 MUs additional endowment) were reduced by 125 MUs. Thus, the punishment condition resembles an ultimatum game (Güth et al., 1982) with the exception that player B has a larger set of available punishment actions.

Player A participated in a sequence of 24 trials in total, facing a different player B in every trial. Control and punishment conditions were randomized across trials. In every trial, there was an initial rest phase of about 6s (Figure 1). Then player A was informed about whether he was in the punishment or the control condition. Player A could then submit his decision ("decide" in Figure 1). After the decision player A had to wait until being informed about player B's punishment decision and the associated final payoffs of the players

- Figure 1 –

In the control condition, a purely self-interested player A should retain the entire endowment for himself. Behavioral studies (Fehr and Fischbacher, 2004; Kahnemann et al., 1986a,b) indicate, however, that the equal split is the most salient fairness norm in situations such as the described control and punishment conditions, as there is simply no reason why A deserves more than B. Evidence for this fairness norm comes, for example, from so-called third party punishment experiments in which a third player (whom we label player C) is informed about how much A gives to B (Fehr and Fischbacher 2004). Then the third player

has the opportunity to punish player A. Typically, player C punishes selfish deviations from the equal split and the punishment is the stronger the more player A deviates in the selfish direction from the equal split. Moreover, Fehr and Fischbacher (2004) elicited fairness judgments which clearly indicate that the egalitarian solution is perceived to be the fair solution and that deviations from the equal split are considered unfair. Thus, fairness judgments, and the observed punishment patterns of third parties all indicate that there is a clear fairness norm in situations where player A can share an endowment with player B. However, in anonymous interactions without a punishment opportunity, such as the above control condition, only a minority of the players typically obeys the fairness norm completely. A significant proportion of the subjects does not share at all and another substantial proportion gives a positive amount but less than the equal split to player B (Camerer, 2003). In the punishment condition, however, player A is likely to face a strong incentive for approaching the equal split because he will be punished severely otherwise. One of the key questions, therefore, was how many more MUs player A transfers to B in the punishment condition, and which brain circuits are associated with this behavioral change.

The control condition measures how much of the 100 MUs player A is willing to give *voluntarily* to player B. This amount can be interpreted as a reflection of A's preference for giving, i.e., as a reflection of his *immediate* self-interest. As a consequence, if A gives amount X to player B in the control condition then A is apparently not willing to give more than X, i.e., giving more than X is not in A's immediate self-interest. Therefore, if the punishment threat in the punishment condition induces A to give more than X the punishment threat requires the inhibition of player A's *immediate* self-interest in order to prevent being punished by B. We therefore conjectured that during the decision phase (Figure 1) lateral prefrontal areas such as the dorsolateral prefrontal cortex (DLPFC), or the ventrolateral prefrontal cortex (VLPFC), which have been shown to be reliably involved in cognitive control and the

inhibition of prepotent responses (Aron et al., 2004; Miller and Cohen, 2001; Sanfey et al., 2003), will be more strongly activated in the punishment condition. This conjecture is also suggested by a recent study (Knoch et al. 2006) where subjects had to take a fair action that yields a lower economic payoff than the self-interested one. If the experimenter disrupted subjects' right DLPFC with low-frequency transcranial magnetic stimulation, the selfish action was chosen much more frequently, and much faster, despite the fact that their fairness judgments remained unaffected. Thus, subjects with right DLPFC disruption behaved as if they could no longer resist the temptation to choose the selfish action, suggesting that when self-interest and fairness are in conflict, the selfish action is the prepotent response. In view of the important role of DLPFC in overriding prepotent impulses (Aron et al., 2004; Miller and Cohen, 2001; Sanfey et al., 2003), this brain area might also play a role in our experiment because subjects' prepotent response (i.e., their immediate self-interest) is to violate the fairness norm.

In the punishment condition player A also has to evaluate the sanctioning threat. Several studies suggest lateral orbitofrontal cortex (OFC) involvement in the evaluation of punishing stimuli that may lead to behavioral changes (Kringelbach, 2005; O'Doherty et al., 2001). These findings led to the conjecture that the OFC might be more strongly activated in the punishment condition.

One of the key features of the present paradigm is that the threat of punishment may induce some subjects to share more equally than when the punishment threat is absent. In order to test whether subjects with a higher propensity to respond to the punishment threat have different personality characteristics we measured subjects' Machiavellism – a combination of selfishness and opportunism – with the Machiavelli questionnaire (Christie and Geis, 1970). In this questionnaire the subjects indicate their degree of agreement with statements such as "It's hard to get ahead without cutting corners here and there" and "The

best way to deal with people is to tell them what they want to hear". It seems plausible to conjecture that subjects with a high Machiavelli score behave more selfishly in the control condition than subjects with a low Machiavelli score. Subjects with higher scores might also increase their transfers substantially in the punishment condition because their opportunism could induce them to adjust their behavior strongly when they face the threat of punishment. In order to explore a possible relation between Machiavellian personality characteristics and the brain circuits activated by the threat of punishment we also computed correlations between individuals' Machiavelli score and their brain activations during decisions under threat of punishment minus control.

Results

Behavioral data

Transfer difference

In the control condition, 11 subjects transferred on average less than five MUs, another 11 subjects transferred between 10 and 20 MUs on average, and only 1 subject transferred more than 30 MUs. This contrasts sharply with the punishment condition, where 16 subjects gave about 50 MUs and only 7 subjects gave less than 40 MUs on average. None of the subjects transferred zero MUs in the punishment condition. On average, A gave about 10 MUs to B during control condition, while A transferred about 40 MUs (Figure 2) in the punishment condition. The transfer difference between conditions is highly significant (Wilcoxon signed rank test, $n = 23$, $p < 0.001$).

- Figure 2 –

Machiavellism and Transfer Difference

We found a negative correlation ($r = -0.47$, $p = 0.026$) between subjects' Machiavelli scores and their average transfers in the control condition and a positive correlation ($r = 0.50$, $p = 0.017$) with their behavioral changes (measured in terms of transfer differences between the punishment and the control condition) caused by the threat of punishment. Therefore, Machiavellian subjects earned the highest incomes because they earned most in the control condition and were best at escaping punishment in the punishment condition. The correlation between overall earnings and the Machiavelli score was 0.48 ($p = 0.026$). In the punishment condition, for example, subjects in the highest quartile of the Machiavelli score earned 41% more than subjects in the lowest quartile.

fMRI data

Since we are interested in the neural circuitry associated with norm compliance behavior, all fMRI data presented below relate to the decision-making epoch of the experiment. As predicted, the contrast (averaged across all subjects) between brain activations during decisions under the threat of punishment and during decisions in the control condition was significant (see Figure 3, Table 1) in the dorsolateral prefrontal (DLPFC; BA 9 and BA 46), the ventrolateral prefrontal (VLPFC; BA 10/46), and the anterior orbitolateral prefrontal cortices (OLPFC; BA 47). All activation increases in the prefrontal cortex occurred bilaterally (Figure 3; see also Figure S1 for time courses). In addition, we found a significant contrast in bilateral medial caudate nucleus (Figure 3D).

- Figure 3 -

Combination of behavioral and fMRI data

Transfer difference

When correlating the increase in brain activations during social punishment trials (punishment minus control) with the increase in MUs transferred to player B in these trials relative to the control condition (transfer difference) we observed strong positive correlations at locations summarized in Table 2. Figure 4 shows scatter plots of the strongest correlations in the right caudate nucleus($r = 0.70$, $p < 0.001$), and the right dorsolateral prefrontal cortex ($r = 0.69$, $p < 0.001$).

- Figure 4, Table 2 -

Machiavellism

When correlating individuals' differential brain activations in punishment minus control with individual Machiavelli scores we observed strong positive correlations in the left anterior OFC (BA 11/47; peak voxel at [–32, 44, –8]; $r = 0.71$; $p < 0.001$). This cluster also contains voxels with a positive correlation between individuals' differential brain activation and the increase in norm compliance in the punishment condition, i.e., the transfer difference across conditions (Figure 5). Additionally, Machiavelli scores were strongly correlated with differential brain activation in the right insula (peak voxel at [42, 8, 0]; $r = 0.64$; $p < 0.001$; Figure 6).

- Figures 5 and 6 -

Even at a rather liberal threshold ($p < 0.05$, extent threshold of 102 voxels corresponding to a level of $p = 0.5$ at the cluster level) we did not find correlations between Machiavelli scores and brain activity in the other main areas of activation, i.e., in bilateral DLPFC, VLPFC, and caudate nuclei (Figure 3). Hence, among the main areas of brain activation (depicted in Figure

3), the correlation between individuals' Machiavelli scores and brain activity appears to be specific to the lateral OFC.

Discussion

In order to investigate the neural correlates of norm compliance, we set up an experimental paradigm with two main conditions: a control condition, where player A could freely distribute the initial endowment of 100 monetary units between himself and player B, and a punishment condition, where A knew that B could punish him after being informed of his sharing decision.

Behavioral results indicate that the punishment threat was very effective in inducing subjects to obey the fairness norm. In fact, several of the A subjects who gave zero MUs in the control condition changed their behavior markedly in the punishment condition, making average transfers close to the equal split. This strong behavioral change across conditions was triggered by the severe punishment that subjects in the role of player B imposed on norm violators – the more player A fell short of the fair transfer level of 50 MUs, the more player B punished him.

As the punishment condition forces subjects to move closer to the fairness norm, we conjectured that forced norm compliance might activate prefrontal brain regions implicated in the evaluation of punishing stimuli (Kringelbach, 2005; O'Doherty et al., 2001) and the inhibition of prepotent responses (Aron et al., 2004; Miller and Cohen, 2001; Sanfey et al., 2003). The fMRI data are consistent with these conjectures (Figure 3). In addition, activations in the dorsolateral and orbitolateral prefrontal cortices were positively correlated with the increase in norm compliance that the punishment threat induces. Subjects with stronger differential activations of the DLPFC exhibited larger transfer increases across conditions. Likewise, subjects with greater differential activation of the orbitolateral prefrontal cortex

showed a stronger increase in norm compliance in the punishment condition. If activation in the lateral OFC represents the punishing stimulus in our task, the strong correlation between lateral OFC activation and increases in norm compliance may indicate that subjects with a stronger subjective representation of the punishment threat show stronger norm compliance in response to this threat.

In addition, we found significantly higher activation in the bilateral medial caudate nucleus (Figure 3D) which plays a crucial role in processing information about positive and negative reinforcers. Single-neuron recording in non-human primates (Schultz, 2000) and neuroimaging studies with humans using money as a reward (Delgado et al., 2003; Knutson et al., 2000, 2001) reliably show caudate activation in tasks involving uncertain rewards or punishments. Moreover, the caudate is primarily activated upon contingency between an action and a reward (O'Doherty, 2004; Schultz et al., 2003; Tricomi et al., 2004), e.g., if a subject can increase his chances of earning additional money or decrease the probability of losing money by taking appropriate actions. Since the monetary reward values of different transfer levels are completely certain in the control condition but highly uncertain in the punishment condition, the caudate is likely to be more strongly activated in the punishment condition. Consequently, higher caudate activity in the punishment condition may represent the expected, yet uncertain, punishment reduction associated with transfer increases. This interpretation is further supported by the fact that the strength of the caudate activation predicts the increase in norm compliance across treatment conditions because subjects who expect a higher punishment reduction from a given transfer increase have reason to increase their transfers more strongly in the punishment condition.

Some researchers (Knutson et al. 2000; 2001) have also pointed out that caudate activity in monetary incentive delay tasks may represent the arousal associated with perceived (but uncertain) reward and punishment levels. This interpretation is different, yet compatible,

with that given above because it makes the same predictions. If caudate activation represents the arousal associated with the punishment threat, one should also observe more caudate activation in the punishment condition. In addition, subjects experiencing more arousal (higher caudate activation) can be plausibly expected to exhibit a stronger increase in norm compliance in the punishment condition.

In addition, we found that Machiavellian personality characteristics played an important role in forced norm compliance and associated brain activity. The subjects' Machiavelli score is strongly correlated with the increase in norm compliance in the punishment condition and with activations in brain regions associated with the evaluation of punishing stimuli (lateral OFC, Figure 5) and the representation of emotional states (insula, see Figure 6). The strong correlations in the left lateral OFC may be interpreted as additional support for the hypothesis that this brain area plays a special role in detecting and evaluating the punishment threat. By definition, Machiavellian subjects are both self-interested and opportunistic, making good abilities in detecting and evaluating threats to their self-interest necessary.

The strong correlation of brain activation in the right insula with individuals' Machiavelli Score is interesting because a number of findings indicate the insula's involvement in emotional experiences such as anger, fear, pain, sadness, and disgust (Craig, 2002). There is also increasing evidence that the insula is a key component of individuals' interoceptive subjective awareness of their bodily states, including states of emotional arousal (Craig, 2002). A recent study involving an introspective task even documented a strong positive correlation between activity in the right insula on the one hand and individual anxiety and negative affect on the other hand (Critchley et al., 2004). In view of these results, Machiavellian subjects may have experienced stronger negative affect during the punishment condition, which may have contributed to their stronger behavioral response in this condition.

An interesting question in the context of this paper is which of the brain activations

observed in Figure 3 and Table 1 are specific to the enforcement of norms prevailing in *social* interactions, and which activations are merely generated by the fact that lower transfer levels trigger a subsequent payoff reduction (punishment). In principle, the evaluation of punishing stimuli or the inhibition of prepotent responses could also play a role if player A is not involved in a social interaction with B but merely faces a punishment threat that a preprogrammed computer executes. Thus, the same brain regions may be implicated in the absence of any kind of social interaction with another human being. It is interesting in this context that the lateral OFC and DLPFC (in particular the right DLPFC) have often been implicated in social moral judgment tasks (Greene et al., 2004; Moll et al., 2003; Moll et al., 2005; Prehn and Heekeren, in press). Perhaps, the stronger activation of these regions in the punishment condition could reflect recruitment in the service of moral and social cognition. However, these alternative interpretations cannot be further explored with the current design.

For this reason we conducted an additional experiment with a "nonsocial" punishment condition, where we kept everything constant relative to the previously described social punishment condition except for the fact that player A did not face a human player B but only a preprogrammed computer. In particular, the computer's punishment "choices" in the nonsocial punishment condition were distributed in the same way as the punishment choices of human players B in the social condition. Thus, player A faced exactly the same probability and size of punishment for any given transfer level in the nonsocial condition as in the social punishment condition, with the difference that a preprogrammed computer – and not a human player – executed the punishment.

The nonsocial punishment condition enables us to answer the question whether the recruitment of the brain regions described in Figure 3 are stronger in the social than the nonsocial condition. We find that, among the areas reported in Figure 3, activation in the right DLPFC (BA46 [28 26 24]), the left VLPFC (BA10 [-42 52 -2]), and in the right OLPFC

(BA47 [44 42 -6]), are higher in the social condition (at $p < 0.005$, Table S1), supporting the view that a punishment threat arising from a social interaction recruits these areas differently. In addition, significantly higher activation in the social context occurred in the right insula, in the left OLPFC (in a different cluster from the one in Figure 3) and in the left superior temporal gyrus (Table S2). Taken together these results indicate that the presence of a social context either activates specific brain areas or modulates activations observed in right DLPFC, left VLPFC, and bilateral OLPFC.

In this study, we sought to uncover the neural circuits involved in forced norm compliance. This question touches the very foundations of human sociality because the establishment of large-scale cooperation through social norms is a unique feature of the human species. Norm compliance among humans is either based on people's voluntary compliance with standards of behavior that are viewed as normatively legitimate or on the enforcement of compliance through punishment. Although much compliance is voluntary, there can be little doubt that social order would quickly break down in the absence of punishment threats because a minority of non-compliers can trigger a process that leads to widespread non-compliance due to the conditional nature of many people's compliance (Fischbacher et al., 2001; Fehr and Gächter 2002).

This is the first study to our knowledge that examines the brain processes involved in humans' *behavioral* response to the threat of punishment for norm violations. We developed a paradigmatic task that enables the study of norm compliance at both the behavioral and brain levels. We therefore believe that our study may contribute to a deeper understanding of important neurobiological aspects of human sociality such as the neurobiological basis of individual differences in norm compliance in healthy subjects which – in view of the strong role of Machiavellian personality traits in our task – includes exciting potential insights into the role of the "Machiavellian Mind" for norm compliance. One possible interpretation of the

fact that Machiavellian subjects earn more money in the social experiment is that they have better reward learning abilities with associated modulation of structures including the OFC and the basal ganglia. Although better reward learning skills may represent a basic neurobiological feature of Machiavellian personalities, it is unlikely that such a multi-faceted trait can be reduced to one single dimension. If Machiavellianism were reducible to reward learning ability one would expect a correlation between Machiavelli score and striatal activations but we fail to observe a correlation between Machiavelli scores and caudate activity. Instead, we find a strong positive correlation between Machiavelli scores and activation in left OLPFC and the right insula; parts of these brain regions are also more strongly recruited in the social punishment condition than in the nonsocial condition. Therefore, Machiavellian personality traits may also rely on abilities and brain regions that are more specifically tied to the social nature of human interactions.

The approach taken in this paper may also illuminate potential defects in the neural circuitry associated with severely diminished norm compliance in people with antisocial personality disorder (Raine and Yang 2006). Work by Veit et al. (2002) and Birbaumer et al. (2005) has, for example, shown that healthy control subjects show significant activation in the lateral OFC and the insula in a fear conditioning paradigm in which physical pain is associated with the conditioned stimulus. In sharp contrast, criminal psychopaths exhibit no such activations in this task. The authors interpret these findings as support for a lack of emotional anticipation of aversive events in criminal psychopaths. The fact that criminal psychopaths show a lack of activation during the acquisition of fear in lateral OFC where we observe correlations with both subjects' Machiavelli scores and their increases in norm compliance under the threat of punishment is very interesting, and suggests exciting hypotheses about psychopaths' behavior and brain activations in our norm compliance task. Unlike Machiavellian subjects, however, psychopaths lack the ability to respond adaptively to

punishment threats; they therefore should show less compliance with the fairness norm than healthy subjects, and this diminished norm compliance should be associated with a lack of or a significantly lower activation in the lateral OFC and the insula.

Our results also hint at a possible lateralization of the neural circuitry involved in forced norm compliance, as we find a strong positive correlation between the increase in norm compliance under the threat of punishment and brain activation in the right, but not the left, DLPFC. The lateralization hypothesis is consistent with a recent finding by Knoch et al. (2006) that disruption of the right, but not left, DLPFC with low-frequency transcranial magnetic stimulation is associated with more selfish and less fair-minded behavior. These findings are also interesting in the light of evidence (Damasio 1995) suggesting that patients with right prefrontal lesions are characterized by the inability to behave in normatively appropriate ways, despite the fact that they are keenly aware of the prevailing fairness norm in tasks like ours (Fehr and Fischbacher 2004). Thus, a dysfunction of the right DLPFC or its specific connections may also underlay certain psychopathological disorders characterized by excessively selfish tendencies and a failure to obey basic social norms.

Finally, our finding that the dorsolateral, ventrolateral, and orbitolateral cortices are part of the neural circuitry involved in social norm compliance may have implications for the criminal justice system. As these brain areas are not yet fully developed in children, adolescents, or even young adults (Gogtay, 2004), our results are consistent with the view that these groups may be less able to activate the evaluative and inhibitory neural circuitry necessary for the appropriate processing of punishment threats. Thus, our results may provide support for the view that the criminal justice system (Garland and Glimcher, 2006) should treat children, adolescents, and immature adults differently from adults.

Supplemental Data

In order to examine the extent to which our main activations (reported in Figure 3 and Table 1) are modulated by the social context of the experiment, i.e., by the existence of a player B who interacts with A and can punish A, we conducted an additional (nonsocial) fMRI experiment with a further sample of 23 healthy, right-handed male students (mean age ± SD, 24.8 ± 1.9 years) as players A. In the nonsocial experiments subjects knew that they play against a preprogrammed computer. The computer was programmed to punish on average in the same way as players B did in the social punishment condition, i.e., the punishment choices of the computer were distributed in the same way as the punishment choices of the human players B in the social punishment condition.

Table S1: Testing for the significance of mean activation differences between the social and the nonsocial punishment condition in the brain areas reported in Figure 3 and Table 1.

Anatomical Region	L/R	BA	t(44)	p
DLPFC	L	9	0.056	0.478
	L	46	0.722	0.237
	R	46	2.792	**0.004**
	R	9	1.004	0.160
VLPFC	L	10	2.859	**0.003**
	R	10/46	0.379	0.087
OLPFC	L	47	2.281	0.014
	R	47	2.697	**0.005**
Caudate Nucleus	L		2.374	0.011
	R		2.084	0.022
Thalamus	L		1.593	0.059
Lingual/Fusiform Gyrus	L	17/18/37	1.097	0.139
Cerebellum	R		1.464	0.075

Notation in Table S2: L/R: left/right; BA: Brodmann´s areas; p: P-values refer to the t-statistics [t(44)] for the null hypothesis that the neural activations between the social and the nonsocial punishment condition are identical. Anatomical labeling and tabulation of Brodmann's areas refer to the main clusters of significant activations in the contrast of punishment minus control in Table 1 of the main text. To compute t-statistics, the activations in the social and the nonsocial punishment trials were averaged over voxels and subjects. Given the null hypothesis of no difference between both conditions, one-sided t-tests (with 44 degrees of freedom) were used to test if mean pooled activation was higher in the social experiment at a significance level of $p < 0.005$. We chose this level because

we restricted our attention to those clusters that had already demonstrated significantly different activations during the social experiment (Punishment-Control). P-values in bold denote significant differences.

Table S2: Significant differences in neural activations in the contrast between the social and the nonsocial punishment condition *in additional brain regions* (i.e., exclusively masked by regions referred to in Table S1).

Anatomical region	**L/R**	**BA**	**x**	**y**	**z**	**z-score**
OLPFC	L	47	-34	46	-6	4.25
Superior Temporal Gyrus	L	38	-26	6	-28	3.90
Superior Temporal Gyrus	L	22	-56	-26	4	3.70
Insula	R		50	2	6	3.48
	R		42	-2	2	3.29

L/R: left/right; BA: Brodmann´s areas ; x, y, z: stereotaxic coordinates (MNI-space); z-thresholds: Z=3.09, p = 0.001. Note that the reverse contrast (Nonsocial-Social) did not reveal any significant differences, even when lowering the significance level to $p < 0.01$.

Table S3: Average total payoff of player A for different transfers to player B

Transfer to player B	0	10	20	30	40	50	60	70	80	90	100
A's average total payoff in the control condition	125	115	105	95	85	75	65	55	45	35	25
A's average total payoff in the punishment condition	8.5	25	15.05	19.25	40.75	53.35	65	55	45	35	25

Note: In the punishment condition A's average total payoff consists of what he keeps from the potential transfer amount of 100 MU minus the punishment that player B (in the social punishment condition) or the pre-programmed computer (in the nonsocial punishment condition) imposes on A.

Figure S1: Mean percent BOLD signal change for control and punishment trials over 23 subjects and difference curves. For each subject, time series were extracted at the peak voxel of clusters of significant differential activation in the contrast punishment minus control (Table 1) and processed in the same way they had entered individual GLM estimation. Then, separate averaging was performed, over 12 punishment and 12 control trials, of time segments in an interval from -5 to +10 seconds (in units of TR) around the point in time when subjects sent their decision, marked "Return decision".Individual mean percent signal change curves for punishment and control were obtained by subtracting, from the respective averaged time segments, the constant term from the GLM, dividing the result by this constant term and then multiplying by 100 percent. Individual difference curves were determined as difference of individual mean percent signal change curves for punishment and control. Finally, condition and difference curves were averaged across subjects, with error bars indicating the standard error of the mean differences.

Figure S2: Significant differences ($p < 0.001$) of neural activations in the contrast between the social and the nonsocial punishment condition in brain regions outside the clusters of Table 1. **(A)** left lateral orbitofrontal cortex (BA 47; x = -34, y = 46, z = -6; z-score = 4.25). It is worth noting that voxels of the corresponding cluster intersect with those voxels bearing significant positive correlations with Machiavellism (see Figure 5 of main text). **(B)** left lateral superior temporal gyrus (Brodmann's area (BA) 38; x = -26, y = 6, z = -28; z-score = 3.90; BA 22; x = -56, y = -26, z = 4; z-score = 3.70). **(C)** right insula (x = 42, y = -2, z = 2; z-score = 3.29). In the whole brain, the reverse contrast did not reveal any significant differences, even when lowering the significance level to $p < 0.01$.

Experimental Procedures

Participants

24 healthy, right-handed male students (mean age ± SD, 23.5 ± 2.3 years) participated as players A in the social fMRI experiment. One of the subjects obviously failed to understand the game and was excluded from further analysis. We therefore refer to a sample of 23 subjects throughout the manuscript.

In order to generate a credible punishment threat, the number of monetary units that could be spent on punishment and the effect of punishment were calibrated in a previous behavioral pilot study. At the end of this study, the subjects in the role of player B agreed that their decisions could be reused in other sessions of the experiment. In the subsequent fMRI study, the scanned players A faced the decisions of those players B and hence faced decisions of real human opponents. Both players A and B were paid real money. However, player B earned money twice, once in the pilot experiment and once in the fMRI experiment in which his decision had been reused.

In 24 trials of the game in the fMRI study, player A faced a player B who could punish him (punishment condition) in half of the trials, while player B was just a passive recipient of A's transfer in the remaining control trials. Prior to scanning, subjects read written instructions describing the sequence of events, the treatment conditions, and the payoff rules. Subjects received a lump sum payment of € 20 for participating plus 1 Euro-Cent for every monetary unit (MU) earned. Only subjects with neither an acute medical condition nor a history of psychiatric or neurological illness could participate in the study. Subjects were scanned only after contraindications had been excluded (through survey of subjects) and subjects had given written informed consent. All methods and procedures used in this study were approved by the human subjects committee of the University of Ulm, Germany, and conform to the Code of Ethics of the World Medical Association (Declaration of Helsinki).

fMRI Paradigm

Player A was randomly matched with another player B in each trial. Player A saw, via fMRI-compatible video goggles mounted above the subject's eyes (Resonance Technology Inc., Northridge, CA, USA), a series of computer screens (Figure 1). In 24 rounds preceding the game, subjects familiarized themselves with the response device, a trackball-operated slider, by matching different predetermined transfers. During the game, transfer decisions were made within 4.55s (standard deviation 3.16s) on average, and 96% of the decisions were made within 10s. Screens were presented, and behavioral and timing data were collected using the software package zTREE (Fischbacher, 2007), a program for conducting behavioral experiments in combination with neuroimaging.

Prior to scanning of each experiment, subjects filled out several questionnaires, including the Machiavelli questionnaire (Christie and Geis, 1974), which consists of two subscales. One measures how Machivellian a subject is, the other measures how Machiavellian a subject expects other people to be. Here we used the former subscale to explore how a subject's degree of Machivellism is related to his brain activity and behavior.

fMRI Acquisition

One of our hypotheses concerned the activation of the lateral orbitofrontal cortex in the punishment condition. Therefore, special care was taken in setting up an imaging protocol that could overcome the increased risk of image artefacts and signal dropouts in this region during BOLD gradient-echo-planar imaging (3 Tesla Siemens Magnetom Allegra, Siemens, Erlangen, Germany). In accordance with previous recommendations (Kringelbach and Rolls, 2004; Schmitz et al. 2005) we used an in-plane matrix resolution of 128x128 pixels (voxel size: 2 x 2 mm^2) and a slice thickness of 2 mm (+ 0.5 mm gap) when scanning 32 transversal slices with repetition time (TR) 2490 ms, echo time 38 ms, and receiver bandwidth 2790 Hz/Pixel in ascending

direction. In order to further reduce possible image artefacts, slices were oriented steeper than AC-PC orientation at an angle of -25° between transversal and coronal planes. Anatomic imaging included a full-brain EPI (parameters as above, 56 slices), and a sagittal MPRAGE T1 acquisition with voxel size 1x1x1 mm^3 and matrix 256x256 pixels.

fMRI Analysis

fMRI data underwent a standardized preprocessing (temporal and spatial realignment, spatial normalization), and general linear model (GLM) analysis with SPM5 (Release 748, http://www.fil.ion.ucl.ac.uk/spm). In each subject, the full-brain EPI was co-registered onto the mean volume of the spatially realigned EPI time series. Then, the T1 image was co-registered onto the full-brain EPI and normalized to the standard MNI T1 template. The resulting transformation procedure was applied to the EPI time series which was resliced to a voxel size of 2x2x2 mm^3. Finally, images were spatially smoothed with a 10 mm full width at half maximum (FWHM) Gaussian kernel.

In order to load signal variance onto all known sources of variance, the GLM contained regressors for the treatment, decision, wait, and informed epochs (see sequence of screens in Figure 1). All regressors except the one for decision were modeled as epochs of 6s length, with onsets at the times of appearance of the respective screens (Figure 1). Epoch length in the decision regressor was 5s. Onset times were determined by counting back 5s from the point in time when the subject had sent his decision, thereby reducing the length of a treatment epoch in case the subject made the subsequent decision within less than 5s. All regressors came in two main conditions, punishment and control, leading to 8 regressors per subject. All regressors were convolved with the canonical hemodynamic response function. Linear contrasts between decision in trials with punishment versus decision in trials without punishment were subjected to a random effects analysis to compute main effects (one-sample t-test), and to regression analyses with behavioral and questionnaire data as regressors. Anatomic labeling of activated regions was done by two

observers (M.S. and G.G.) with experience in neuroanatomy, using an anatomic atlas (Duvernoy, 1999) as well as computationally with the WFU PickAtlas (V1.02, Wake Forest University School of Medicine) and MRIcro (V1.39, Build 4, www.mricro.com).

Acknowledgements

We thank Jo Grothe for his very supportive technical assistance.

This paper is part of the research priority program at the University of Zurich on the "Foundations of Human Social Behavior – Altruism versus Egoism". E.F. and U.F. also gratefully acknowledge support from the National Competence Center for Research (NCCR) in Affective Sciences. The NCCR are financed by the Swiss National Science Foundation.

References

Adenaes, J. (1974). Punishment and Deterrence (Michigan: University of Michigan Press).

Adolphs, R. (2001). The neurobiology of social cognition. Current Opinion in Neurobiology *11*, 231-239.

Adolphs, R. (2003). Cognitive neuroscience of human social behaviour. Nature Reviews Neuroscience *4*, 165-178.

Andreoni, J., Harbaugh, W., and Vesterlund, L. (2003). The Carrot or the Stick: Rewards, Punishments, and Cooperation. American Economic Review *93*, 893-902.

Aron, A. R., Robbins, T. W., and Poldrack, R. A. (2004). Inhibition and the right inferior frontal cortex. Trends in Cognitive Sciences *8*, 170-177.

Birbaumer, N., Veit, R., Lotze, M., Erb, M., Hermann, C., Grodd, W., and Flor, H. (2005). Deficient fear conditioning in psychopathy: a functional magnetic resonance imaging study. Archives of General Psychiatry *62*, 799-805

Boyd, R., Gintis, H., Bowles, S., and Richerson, P. J. (2003). The evolution of altruistic punishment. Proceedings of the National Academy of Sciences of the United States of America *100*, 3531-3535.

Camerer, C. F. (2003). Behavioral Game Theory - Experiments in Strategic Interaction (Princeton, New Jersey: Princeton University Press).

Christie, R., and Geis, F. (1970). Studies in Machiavellism (New York, Academic Press).

Craig, A. D. (2002). How do you feel? Interoception: the sense of the physiological condition of the body. Nature Review Neurosciences *3*, 655-666.

Critchley, H. D., Wiens, S., Rotshtein, P., Ohman, A., and Dolan, R. J. (2004). Neural systems supporting interoceptive awareness. Nature Neuroscience *7*, 189-195.

Damasio, A. R. (1995). Descartes' Error: Emotion, Reason and the Human Brain (New York, Hayrer Collins).

De Quervain, D. J. F., Fischbacher, U., Treyer, V., Schelthammer, M., Schnyder, U., Buck, A., and Fehr, E. (2004). The neural basis of altruistic punishment. Science *305*, 1254-1258.

Delgado, M. R., Frank, R. H., and Phelps, E. A. (2005). Perceptions of moral character modulate the neural systems of reward during the trust game. Nature Neuroscience *8*, 1611-1618.

Delgado, M. R., Locke, H. M., Stenger, V. A., and Fiez, J. A. (2003). Dorsal striatum responses to reward and punishment: effects of valence and magnitude manipulations. Cognitive Affective Behavioral Neuroscience *3*, 27-38.

Duvernoy, H. M. (1999). The human brain. Surface, blood supply, and three-dimensional anatomy, 2nd edn (Vienna: Springer).

Elster, J. (1989). The Cement of Society - A Study of Social Order (Cambridge: Cambridge University Press).

Fehr, E., and Fischbacher, U. (2004). Third-party punishment and social norms. Evolution and Human Behavior *25*, 63-87.

Fehr, E., and Gächter, S. (2002). Altruistic punishment in humans. Nature *415*, 137-140.

Fischbacher, U. (2007). z-Tree: Zurich toolbox for ready-made economic experiments. Experimental Economics *10*, 171-178.

Fischbacher, U., Gächter, S., and Fehr, E. (2001). Are people conditionally cooperative? Evidence from a public goods experiment. Economics Letters *71*, 397-404.

Frith, U., and Frith, C.D. (2003). Development and neurophysiology of mentalizing. Philos Trans R Soc Lond B Biol Sci *358*, 459-473.

Garland, B. and Glimcher, P. W. (2006). Cognitive neuroscience and the law. Current opinion in neurobiology *16*, 130-134.

Glimcher, P. W., Dorris, M. C., and Bayer, H. M. (2005). Physiological utility theory and the neuroeconomics of choice. *52*, 213-256.

Gogtay, N., Giedd, J. N., Lusk, L., Hayashi, K. M., Greenstein, D., Vaituzis, A. C., Nugent, T. F., 3rd, Herman, D. H., Clasen, L. S., Toga, A. W., *et al.* (2004). Dynamic mapping of human cortical development during childhood through early adulthood. Proceedings of the National Academy of Sciences of the United States of America *101*, 8174-8179.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., and Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. Science *293*, 2105-2108.

Greene, J.D., Nystrom, L.E., Engell, A.D., Darley, J.M., and Cohen, J.D. (2004). The neural bases of cognitive conflict and control in moral judgment. Neuron *44*, 389-400.

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., and Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. Neuron *44*, 389-400.

Güth, W., Schmittberger, R., and Schwarze, B. (1982). An Experimental Analyses of Ultimatium Bargaining. Journal of Economic Behavior and Organization *3*, 367-388.

Hsu, M., Bhatt, M., Adolphs, R., Tranel, D., and Camerer, C. F. (2005). Neural systems responding to degrees of uncertainty in human decision-making. Science *310*, 1680-1683.

Kahneman, D., Knetsch, J. L., and Thaler, R. (1986a). Fairness and the Assumptions of Economics. In Rational choice - the contrast between economis and psychology, R. M. Hogarth, and M. W. Reder, eds. (Chicago: University of Chicago Press), pp. 101-116.

Kahneman, D., Knetsch, J. L., and Thaler, R. (1986b). Fairness as a Constraint on Profit Seeking - Entitlements in the Market. American Economic Review *76*, 728-741.

King-Casas, B., Tomlin, D., Anen, C., Camerer, C. F., Quartz, S. R., and Montague, P. R. (2005). Getting to know you: reputation and trust in a two-person economic exchange. Science *308*, 78-83.

Knoch, D., Pascual-Leone, A., Meyer, K., Treyer, V., and Fehr, E. (2006) Diminishing reciprocal fairness by disrupting the right prefrontal cortex. Science 314, 829-832.

Knutson, B., Adams, C. M., Fong, G. W., and Hommer, D. (2001). Anticipation of increasing monetary reward selectively recruits nucleus accumbens. Journal of Neuroscience *21*, RC159.

Knutson, B., Westdorp, A., Kaiser, E., and Hommer, D. (2000). FMRI visualization of brain activity during a monetary incentive delay task. Neuroimage *12*, 20-27.

Kringelbach, M. L. (2005). The human orbitofrontal cortex: linking reward to hedonic experience. Nature Review Neurosciences *6*, 691-702.

Kringelbach, M. L., and Rolls, E. T. (2004). The functional neuroanatomy of the human orbitofrontal cortex: evidence from neuroimaging and neuropsychology. Progress in Neurobiology *72*, 341-372.

Lieberman, M.D. (2007). Social cognitive neuroscience: a review of core processes. Annu Rev Psychol *58*, 259-289.

Miller, E. K., and Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. Annual Review of Neuroscience *24*, 167-202.

Moll, J., de Oliveira-Souza, R., Eslinger, P. J., Bramati, I. E., Mourao-Miranda, J., Andreiuolo, P. A., and Pessoa, L. (2002). The neural correlates of moral sensitivity: A functional magnetic resonance imaging investigation of basic and moral emotions. Journal of Neuroscience *22*, 2730-2736.

Moll, J., de Oliveira-Souza, R., and Eslinger, P.J. (2003). Morals and the human brain: a working model. Neuroreport *14*, 299-305.

Moll, J., Zahn, R., de Oliveira-Souza, R., Krueger, F., and Grafman, J. (2005). Opinion: the neural basis of human moral cognition. Nature Review Neurosciences *6*, 799-809.

Ochsner, K.N. (2004). Current directions in social cognitive neuroscience. Curr Opin Neurobiol *14*, 254-258.

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., and Dolan, R. J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. Science *304*, 452-454.

O'Doherty, J., Kringelbach, M. L., Rolls, E. T., Hornak, J., and Andrews, C. (2001). Abstract reward and punishment representations in the human orbitofrontal cortex. Nature Neuroscience *4*, 95-102.

Prehn, K., and Heekeren, H.R. (in press). Moral Judgement and the Brain: Interaction of Emotion and Cognition. In The Moral Brain, J. Braeckman, J. Verplaetse, and J. DeSchrijver, eds. (Berlin: Springer).

Raine, A. and Yang, Y. (2006). Neural foundations to moral reasoning and antisocial behavior. Social Cognitive and Affective Neuroscience 1, 203-213.

Rilling, J. K., Gutman, D. A., Zeh, T. R., Pagnoni, G., Berns, G. S., and Kilts, C. D. (2002). A neural basis for social cooperation. Neuron *35*, 395-405.

Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., and Cohen, J. D. (2003). The neural basis of economic decision-making in the ultimatum game. Science *300*, 1755-1758.

Schmitz, B.L., Aschoff, A.J., Hoffmann, M.H., and Gron, G. (2005). Advantages and pitfalls in 3T MR brain imaging: a pictorial review. AJNR American Journal of Neuroradiology *26*, 2229-2237.

Schultz, W. (2000). Multiple reward signals in the brain. Nature Reviews Neuroscience *1*, 199-207.

Schultz, W., Tremblay, L., and Hollerman, J. R. (2003). Changes in behavior-related neuronal activity in the striatum during learning. Trends in Neurosciences *26*, 321-328.

Singer, T., Seymour, B., O'Doherty, J. P., Stephan, K. E., Dolan, R. J., and Frith, C. D. (2006). Empathic neural responses are modulated by the perceived fairness of others. Nature *439*, 466-469.

Sober, E., and Wilson, D. S. (1998). Unto Others - The Evolution and Psychology of Unselfish Behavior (Cambridge, Massachusetts: Harvard University Press).

Tricomi, E. M., Delgado, M. R., and Fiez, J. A. (2004). Modulation of caudate activity by action contingency. Neuron *41*, 281-292.

Veit, R., Flor, H., Erb, M., Hermann, C., Lotze, M., Grodd, W., and Birbaumer, N. (2002). Brain circuits involved in emotional learning in antisocial behavior and social phobia in humans. Neuroscience Letters *328*, 233-236.

Legends

Figure 1: Timeline of screens within a single trial of the fMRI experiment. The first screen ("rest") shows a fixation symbol that indicates the beginning of a new trial; it was displayed for a duration of 6s plus a jitter in the range of one scanner repetition time (TR), drawn from a uniform distribution ("≈6s"). The symbols 0:0 and 5:1 on the second screen ("treatment") indicate the control and punishment condition, respectively. The first number in a symbol represents the reduction of player A's, the second number the reduction of B's income per punishment point. The third screen ("decide") informed player A that he could make his decision, with no time restrictions, thereby introducing a second jitter ("self-paced"). In the punishment condition, the fourth screen ("wait") indicates that player B is making his decision. In the control condition, player A also faced the "wait" screen for the same length of time before profits and losses of both players were displayed in the fifth, "informed" screen.

Figure 2: Player A's average transfer to player B over 12 trials for both the control and punishment conditions. Player A's transfer is much closer to the fairness norm of 50% in the punishment condition, indicating the effectiveness of the punishment threat.

Figure 3: Significantly different activations in the punishment condition relative to the control condition where no punishment of player A was possible. Images are sliced at peak of activation given in Table 1. Transversal and coronal slices are indicated by z- and y-coordinates, respectively. **(A)** Bilateral dorsolateral prefrontal cortex (DLPFC, Brodmann Areas (BA) 46 and 9; sliced at [-32, 26, 28]; x,y,z-coordinates in MNI-space). **(B)** Bilateral ventrolateral prefrontal cortex (VLPFC, BA 10/46; [46, 48, 4]). **(C)** Bilateral orbitolateral prefrontal cortex (OLPFC, BA 47; [44, 42, -6]). **(D)** Bilateral caudate nucleus [16,10,14].

Figure 4: Positive correlations between individual activation differences in the contrast punishment minus control and individual transfer differences between punishment and control conditions. Subjects with higher activation differences exhibit higher transfer increases under the threat of punishment. **(A)** right caudate nucleus, **(B)** right dorsolateral prefrontal cortex. The dashed lines represent the best linear fit; r denotes the correlation coefficient. Correlations are significant at a level of $p < 0.001$ (see also Table 2). Coordinates refer to peak activations and are in mm.

Figure 5: Anatomical overlap of correlations of both transfer differences and Machiavelli scores with differential brain activations from the contrast punishment minus control in left lateral orbitofrontal cortex. The highest correlation with transfer differences ($r = 0.67$; $p < 0.001$; Table 2) occurs at MNI location [-36, 48, -10]. The highest correlation with the Machiavelli score ($r = 0.71$; $p < 0.001$) occurs at [-32,44,-8]. Panel **(A)** shows a sagittal and coronal zoom of this region. Voxels with significant correlations between transfer difference and differential brain activation are color-coded in blue. Red color denotes significant correlation with the Machiavelli sore. Overlapping voxels are in violet. **(B)** and **(C)** show scatterplots of correlations calculated from voxel [-34, 48, -10] in the overlapping region; transfer difference: z-score = 3.28; $p < 0.001$; Machiavelli score: z-score = 3.34, $p < 0.001$, with dashed lines representing the best linear fit; r denotes the correlation coefficient. Coordinates are in mm.

Figure 6: Correlation of Machiavelli scores with differential brain activations from the contrast punishment minus control in the right insula ($p < 0.001$; $x = 42$, $y = 8$, $z = 0$). **(A),** location in MNI space, **(B),** correlation plot. Dashed lines represent the best linear fit; r denotes the correlation coefficient. Coordinates are in mm.

Table 1: Significant differences in brain activation in the contrast punishment minus control

Anatomical Region	L/R	BA	x	y	z	Z score
DLPFC	L	9	-32	26	28	3.69
	L	46	-42	44	20	3.43
	R	46	28	26	24	3.95
	R	9	38	38	28	3.22
VLPFC	L	10	-42	52	-2	3.62
	L	10/46	-34	50	6	3.5
	R	10/46	46	48	6	3.16
OLPFC	L	47	-40	36	-14	2.98
	R	47	44	42	-6	2.94
Caudate Nucleus	L		-14	8	12	3.86
	L		-20	2	20	3.00
	R		16	10	14	5.01
	R		14	-2	20	3.34
Thalamus	L		-6	-4	4	3.51
Fusiform Gyrus	L	18/37	-28	-68	-18	3.91
Lingual Gyrus	L	18	-6	-82	-12	4.11
		17/18	-18	-80	-12	4.02
Cerebellum	R		30	-64	-28	3.47
			2	-60	-24	3.32
			16	-76	-28	3.05

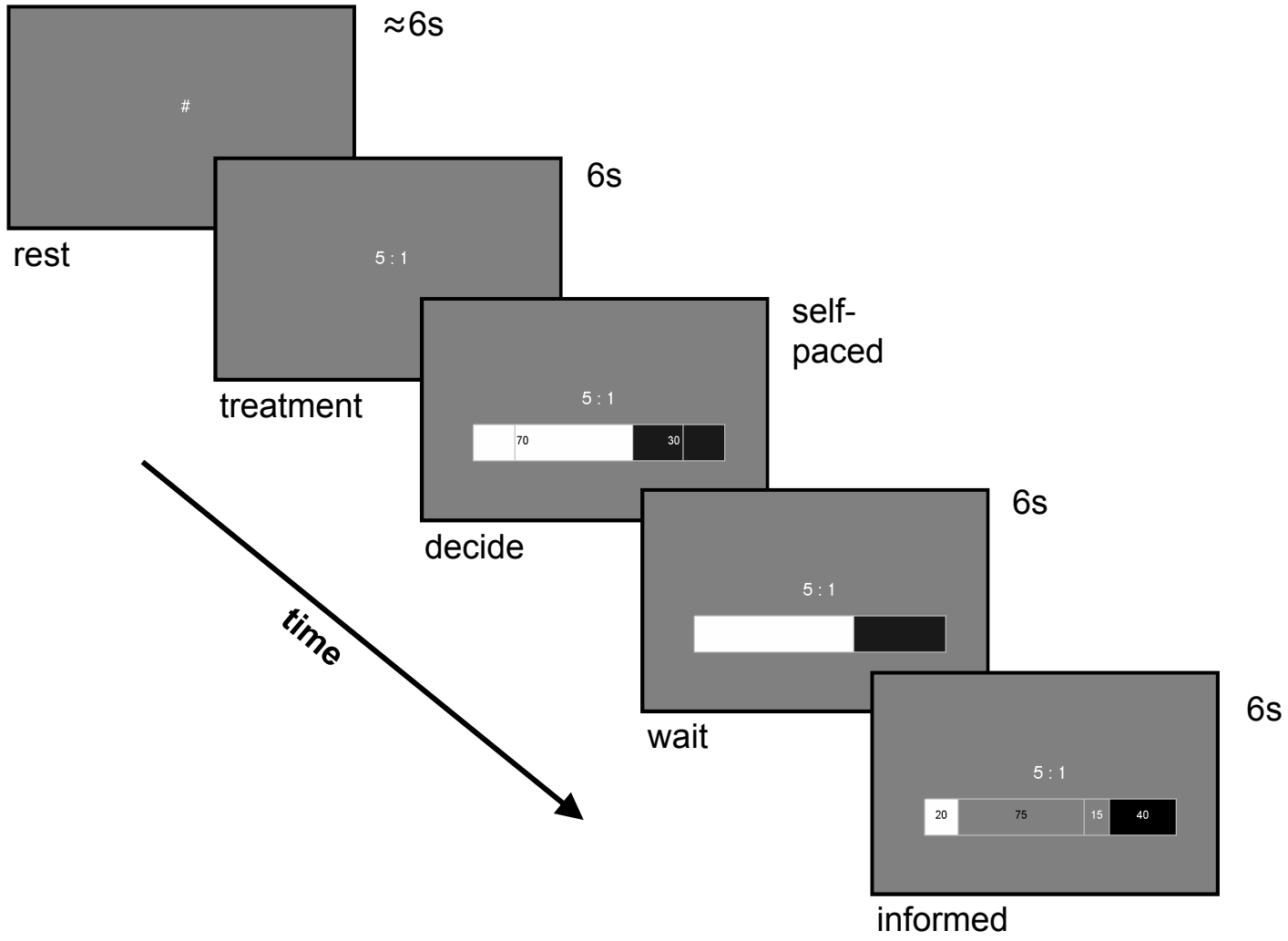
Notation in Table 1: L/R: left/right; BA: Brodmann´s areas; x, y, z: stereotaxic coordinates (MNI-space). z-thresholds and associated significance level: Z=3.09, p = 0.001; Z=2.81, p=0.0025. Predicted activations in italics;

Table 2: Significant correlations between individual transfer differences and individual brain activations in the contrast punishment-control

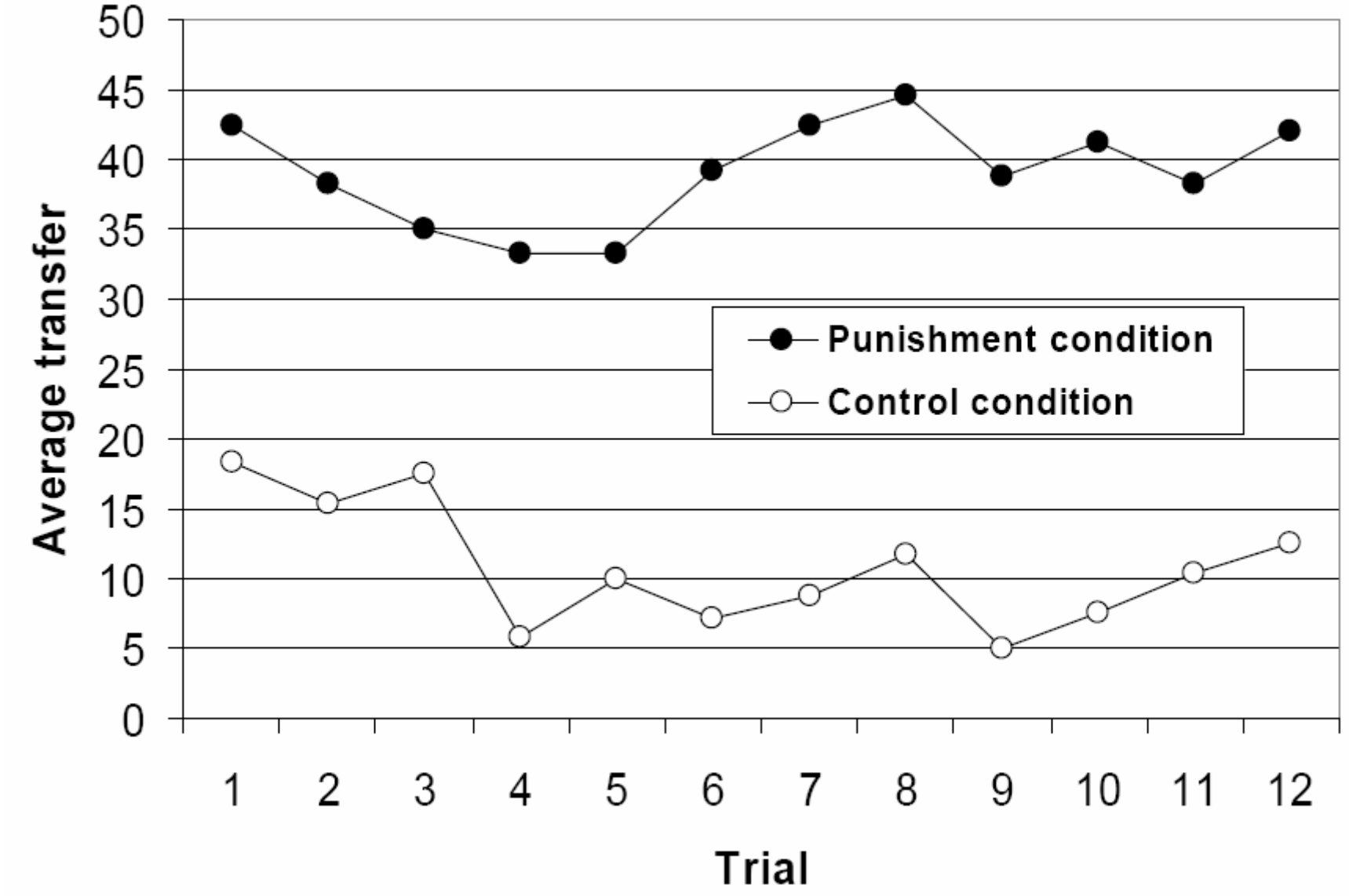
Anatomical region	L/R	BA	x	y	z	z-score	r
Caudate nucleus	R		12	-4	20	3.72	0.700
	L		-20	-10	22	3.42	0.659
DLPFC	R	46	52	28	14	3.67	0.694
OLPFC	L	11	-36	48	-10	3.46	0.665
Inferior frontal gyrus	R	45	60	22	10	3.37	0.652
Middle/Inferior temporal gyrus	R	21/20	56	-34	-6	3.35	0.650
	L	21/20	-56	-34	-16	3.82	0.713

L/R: left/right; BA: Brodmann´s areas; x, y, z: stereotaxic coordinates in mm (MNI-space); Z=3.09, p = 0.001; r: correlation coefficient

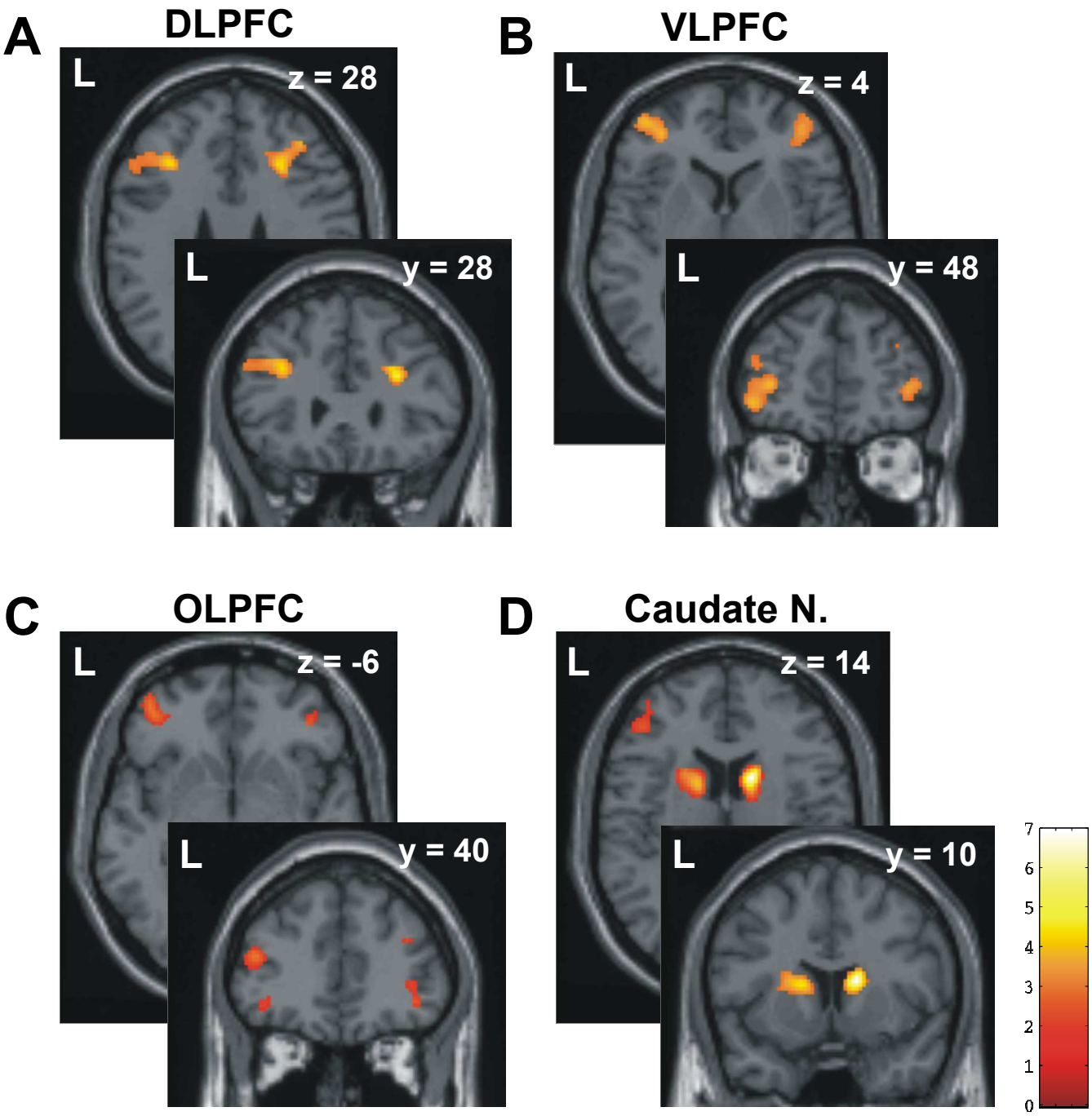
Spitzer et al.
Figure 1



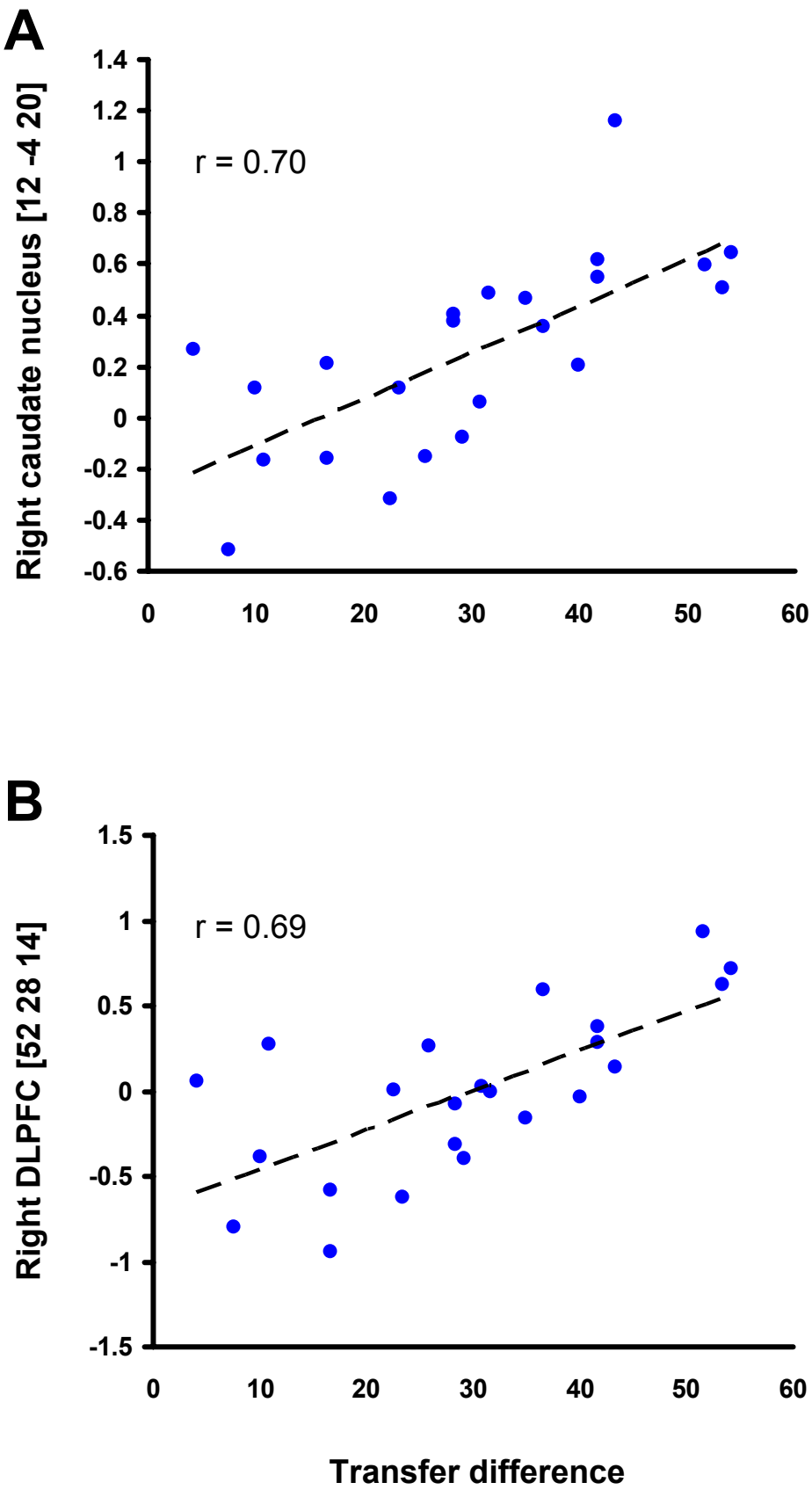
Spitzer et al.
Figure 2



Punishment minus Control

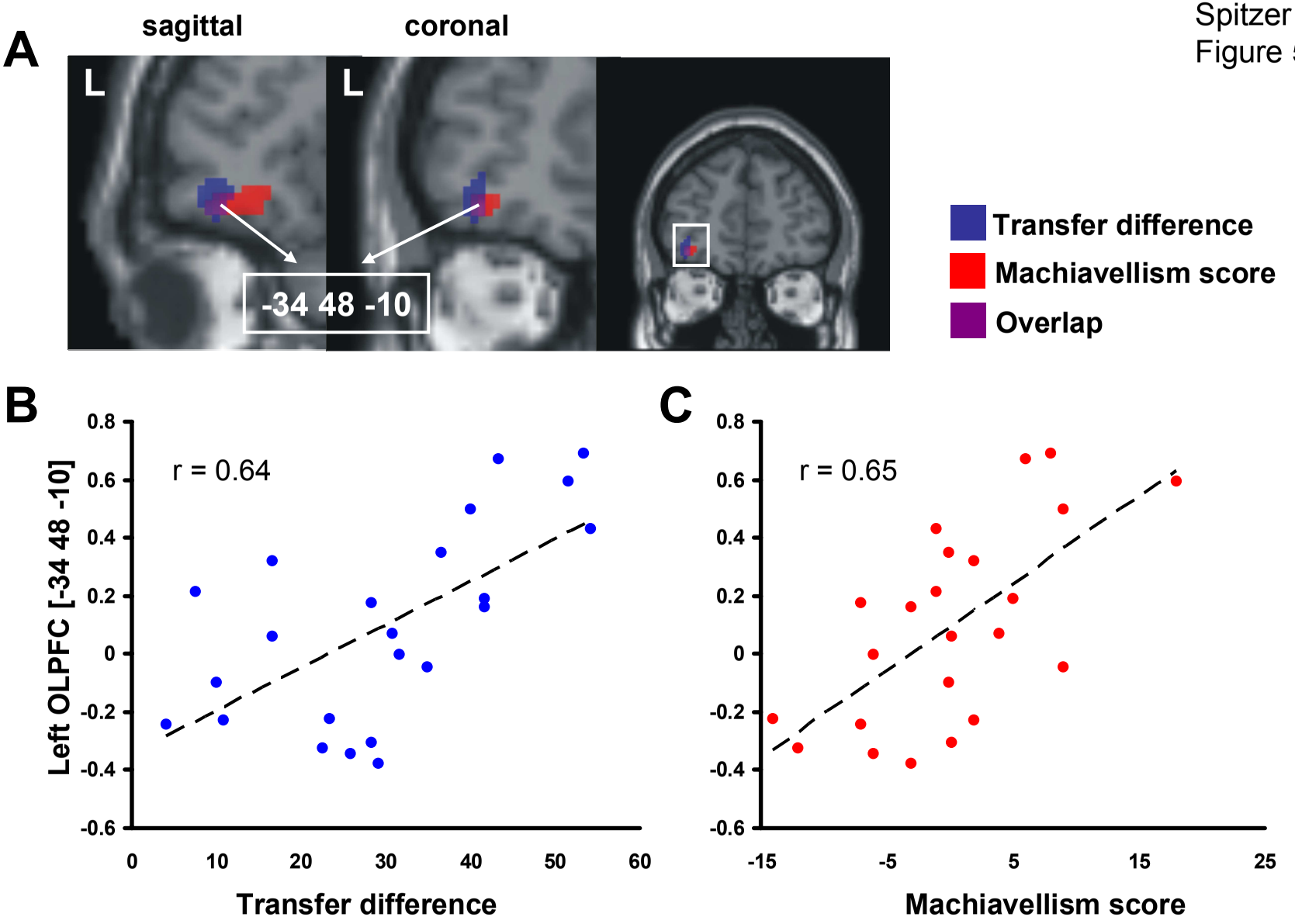


Spitzer et al.
Figure 3

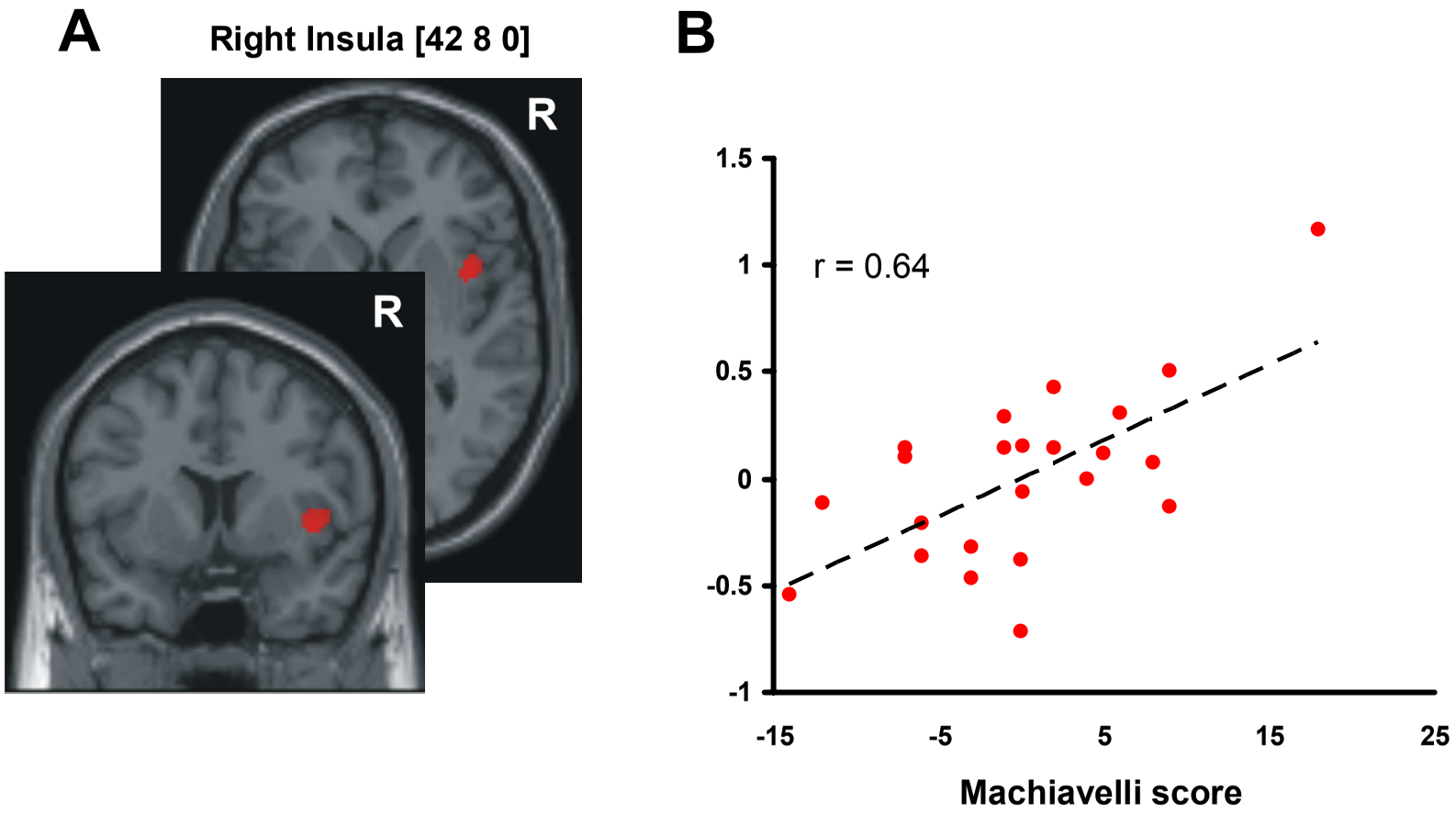


Spitzer et al.
Figure 4

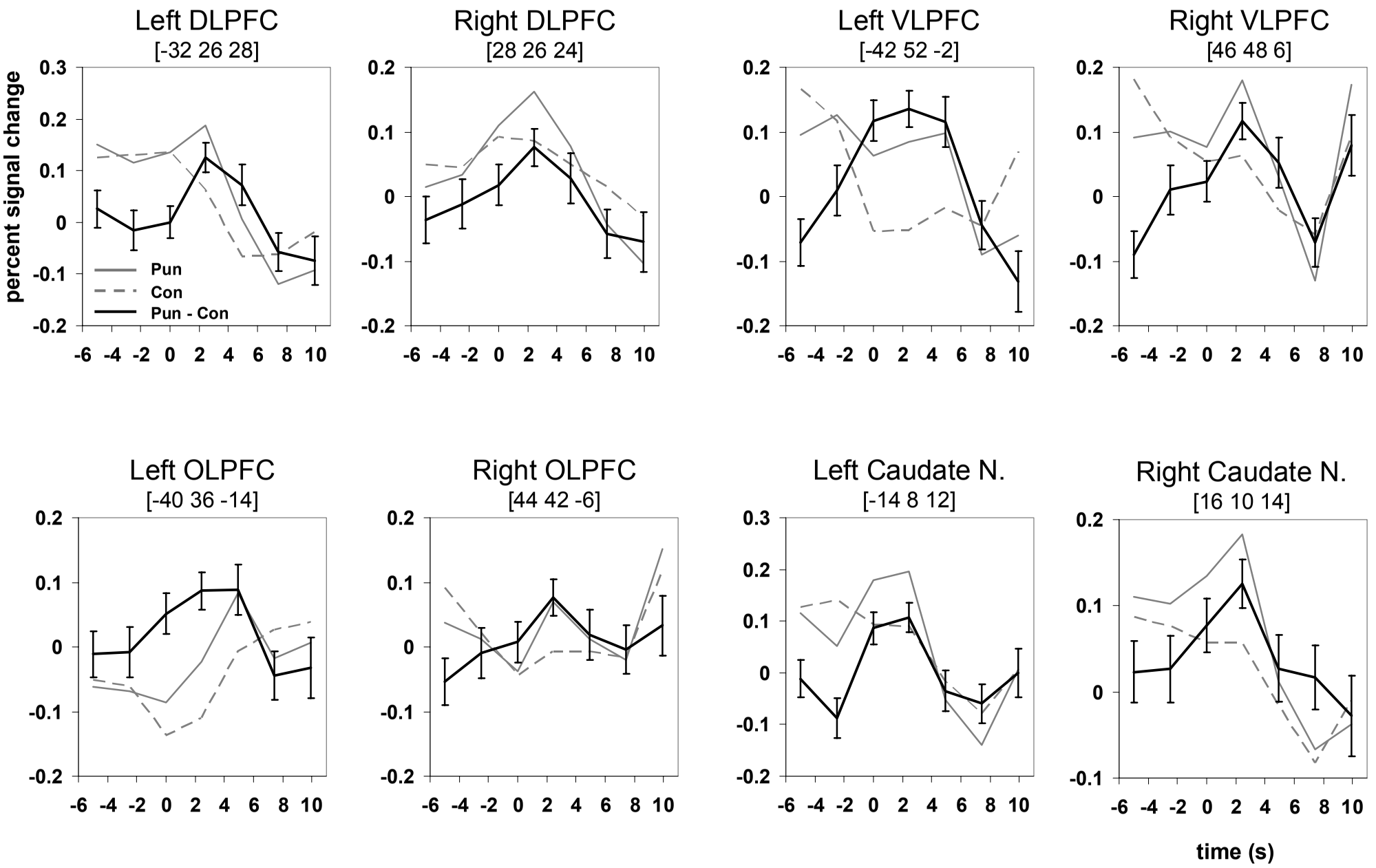
Spitzer et al.
Figure 5



Spitzer et al.
Figure 6



Spitzer et al.
Figure S1



Spitzer et al.
Figure S2

Punishment: Social minus Nonsocial